STOCHASTIC GRADIENT LANGEVIN DYNAMICS FOR BAYESIAN LEARNING

ZEYI XU

Abstract. Stochastic Gradient Langevin Dynamics (SGLD) is a sampling method proposed in the paper *Bayesian Learning via Stochastic Gradient Langevin Dynamics* [3]. It stands at the intersection of stochastic optimization and Langevin-proposed MCMC methods. In this report, we summarize the mathematical foundations of the method, implement the method and test its performance on various learning tasks.

Contents

1. Introduction

The main points of the report are listed as follows:

- The paper considers the Bayesian learning problem, that is, given probability model, do the Maximum A Posterior estimate from sample data.
- In the paper, the authors focus on the Bayesian modeling with parametric models.
- The authors derived the SGLD algorithm by combining Stochastic gradient method with Langevin dynamics.
- In essence, the problem is about sampling from a known distribution. Besides all the standard MCMC methods, Hamiltonian Monte Carlo (HMC) is of interest as an accelerating sampling method. [2] develops a way to incorporate stochastic gradient into HMC method. Recent years, generative models have caught wide attention for their expressive power and scaling property [1].
- In this report, we used SGLD to solve 2 Bayesian learning problems. One is the linear regression problem, the other one is the logistic regression problem.

Date: March 23, 2025.

The rest of the report is organized as follows. In section 2, we briefly go over the problem setting, and elaborate the method as a combination of stochastic gradient method and Langevin dynamics. In section 3, we implement the method and report the results of numerical experiments.

2. Method

We consider the n-dimensional Bayesian learning problem with a large scale dataset. Let $\theta \in \mathbb{R}^n$ be a set of model parameters, $\{x_i\}_{i=1}^N$ data points. The problem is to compute the MAP estimation

$$p(\theta|x) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta) \tag{1}$$

using prior distribution $p(\cdot)$ and likelihood function $p(x|\theta)$. Equivalently, maximize the logarithm of posterior distribution density

$$\max_{\theta} f(\theta) := \log p(\theta) + \sum_{i=1}^{N} \log p(x_i|\theta). \tag{2}$$

For this optimization problem, a classic method is to do gradient ascent,

$$\theta_{k+1} = \theta_k + \alpha_k \nabla f(\theta_k). \tag{3}$$

The drawback of this method is that the model replies deterministically on the training data $\{x_i\}$, and it might overfit and could not generalize to new data. To circumvent this situation, consider adding a stochastic term into the update

$$\theta_{k+1} = \theta_k + \alpha_k \nabla f(\theta_k) + \sqrt{2\alpha_k}\xi_k, \tag{4}$$

where $\xi_k \sim N(0, I_n)$ is a n-dimensional standard normal vector. It can be seen as the α_k-discretization of the following Langevin-dynamics

$$d\theta_t = -\nabla U(\theta_t)dt + \sqrt{2}dB_t, \tag{5}$$

where U is the potential, $U(\theta) = -f(\theta)$. In continuous-time dynamics, (5) converges to its stationary distribution

$$\pi(\theta) \propto e^{-U(\theta)}. \tag{6}$$

In discrete time, such property will not naturally hold, but it can be fit into the framework of Rosenbluth-Hastings by adding a probability of acceptance, making it a typical MCMC algorithm with the proposal in the form of (4). This is called the Metropolis adjusted Langevin algorithm (MALA).

Though provably faster than the naïve proposals such as Gaussian, MALA needs to evaluate the gradients of all current sample points each step. This becomes computationally expensive when (1) the sample size is extremely large (2) the gradient is hard to evaluate.

Stochastic Gradient provides a way to overcome this issue. Instead of evaluating the full gradient, it only computes gradients for a mini-batch of samples $\{x_{ki}\}_{i=1}^m$, and replace the gradient term by the partial gradient on these points. Combining it with (2), (4), we get the so-called Stoachastic Gradient Langevin Dynamics

$$\theta_{k+1} = \theta_k + \alpha_k \left[\log p(\theta_k) + \frac{N}{n} \sum_{i=1}^{n} \log p(x_{ki}|\theta_k) \right] + \sqrt{2\alpha_k}\xi_k \tag{7}$$

where $\xi_k \sim N(0, I_n)$.

3. Numerical Experiments

In this section, we perform 2 numerical experiments to validify SGLD in Bayesian learning tasks. All the necessary code for this section can be found at GitHub repository https://github.com/ZeyiXuu/SGLD.git. As the original paper does not come with code, the code above is adapted from the existing third-party code on SGLD: https://github.com/LouisBouvier/BML_Stochastic_Langevin_Dynamics.git and https://github.com/JavierAntoran/Bayesian-Neural-Networks.git.

3.1. **Bayesian Linear Regression using SGLD.** In this experiment, we apply Stochastic Gradient Langevin Dynamics to perform Bayesian Linear Regression with a Laplace prior of scale $b = 1$. For now, we fix the variance of the Gaussian likelihood to $\sigma^2 = 1$. We also compare Stochastic Gradient Langevin Dynamics to other sampling methods to test its efficiency, particularly its speed.

The model is as follows:

$$p(\beta) \sim \mathcal{L}(0,1), \quad \text{i.e.} \quad p(\beta) \propto \exp(-\|\beta\|_1),$$

For all $i \in \{1, \dots, N\}$,

$$y_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2), \quad \text{i.i.d.}$$

Thus, the log posterior is:

$$\log p(\beta|X, y) \propto -\sum_{i=1}^{N} (x_i^\top \beta - y_i)^2 - \|\beta\|_1 = -\|X\,\beta - y\|_2^2 - \|\beta\|_1$$

Using the fact that a subgradient of the log-prior is $\partial_\beta \log p(\beta) = -\text{sign}(\beta)$ and letting X represent the data matrix, i.e.

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times d}$$

and y the vector of labels, we can write a subgradient of the log-posterior as:

$$\partial_\beta \log p(\beta|X, y) = -X^\top X\,\beta + X^\top y - \text{sign}(\beta).$$

To test validity of the SGLD method, We apply it to synthetic data. Let $y = 0.02x^3 + x^2 + 0.3x$ be the objective relation. Let x be 250 points randomly chosen from $[-5, 5]$, and corresponding y be computed from the desired relation. Apply Mini Batch SGLD method with `batch_size` $= 5$ for 5000 iterations. The resulting parameters form a well-trained linear regression model. Use the model to do prediction, the results are quite close to the ground truth (Fig. 1).

3.2. **Bayesian Logistic Regression using SGLD.** In this experiment, we apply Stochastic Gradient Langevin Dynamics to perform Bayesian Linear Regression with a Laplace prior of scale $b = 1$. Notice that, in this problem, the log posterior becomes:

$$\log p(\beta|X, y) \propto \sum_{i=1}^{N} \log \sigma(y_i \cdot \beta^\top x_i) - \|\beta\|_1$$

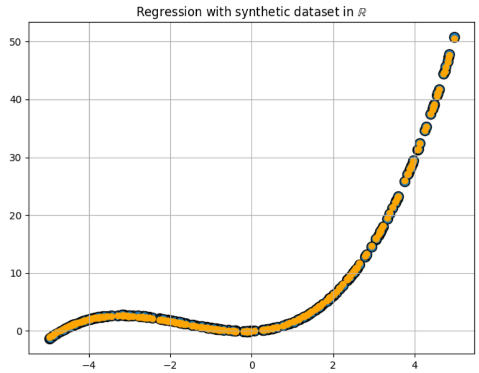

Figure 1. *The performance of SGLD on linear regression tasks*

where σ is the sigmoid function: $\sigma(x) = \frac{1}{1+\exp(-z)}$. We can then write a subgradient of the log-posterior using the fact that $\forall z \in \mathbb{R}, \sigma'(z) = (1 - \sigma(z)) \cdot \sigma(z)$, and the subgradient of the log-prior is $\partial_\beta \log p(\beta) = -\text{sign}(\beta)$:

$$\partial_\beta \log p(\beta|X, y) = \sum_{i=1}^{N} y_i \cdot x_i \left(1 - \sigma(y_i \cdot \beta^\top x_i)\right) - \text{sign}(\beta)$$

To test validity of the SGLD method, We apply it to synthetic data. We generate two clusters containing 1500 sample points centered respectively at $(-2, 2)$ and $(2, -2)$ with `std` $= 1.3$. Apply Mini Batch SGLD method with `batch_size` $= 5$ for 5000 iterations. The resulting parameters form a well-trained Logistic regression model. Use the model to do prediction, we gained a high accuracy for point separation (Fig. 2(a)). Fig 2(b) is the contour map of the predicted probability given by the regression model. When a point deviates from the straight line $y = -x$, its predicted value will rise or drop quickly to 1 or 0, showing that the model will give it a clear and correct classification.

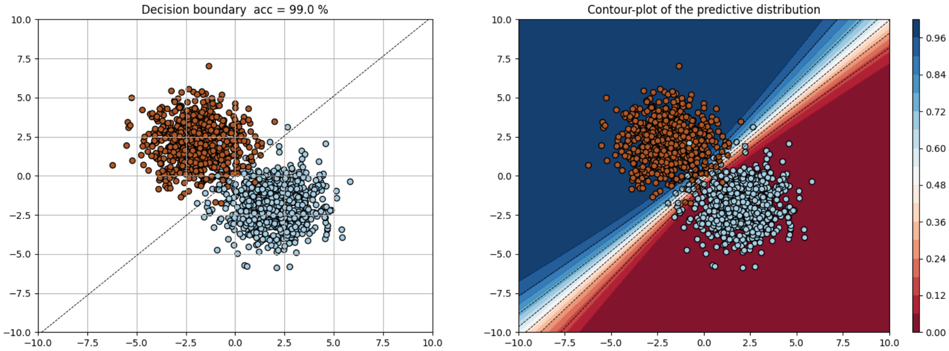

Figure 2. *Classification results of the Bayesian Logistic Regression model*

4. Conclusion

In this project, we summarize the mathematical aspects of the Stochastic Gradient Langevin Dynamics (SGLD) as a combination of Stochastic Gradient Descent with Langevin Dynamics. We implement the algorithm and test it on simple regression tasks with synthetic data. The results show the validity of this method. Future directions of this project are:

(1) comparing the efficiency of SGLD with other more naive sampling methods;
(2) recognizing the accelerated methods to SGLD, such as SAGA-LD and SVRG-LD, and developing more efficient Bayesian learning methods;
(3) using acclerated optimization methods from math community to further accelerate Langevin-based sampling methods.

References

[1] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, Nov. 2022. 1

[2] T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo, 2014. 1

[3] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011. Omnipress. 1